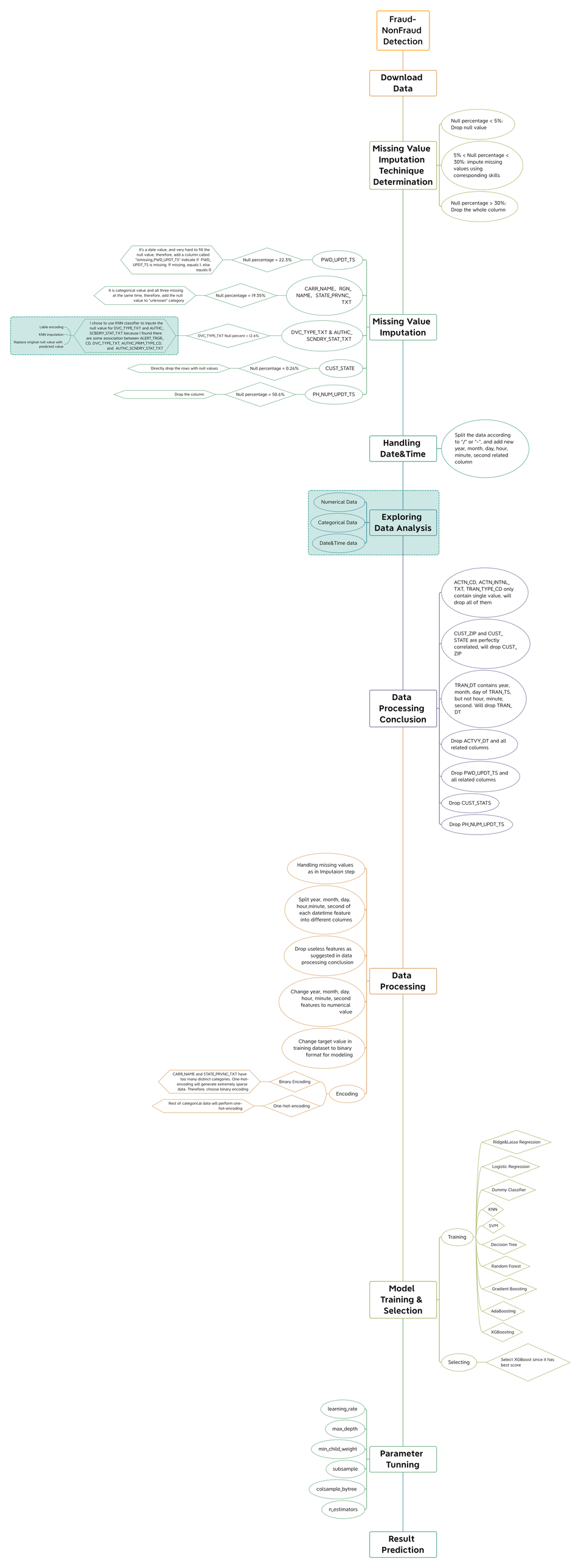
Fraud-NonFraud Detection
Download Data
Missing Value Imputation Techinique Determination
Null percentage < 5%: Drop null value
5% < Null percentage < 30%: Impute missing values using corresponding skills
Null percentage > 30%: Drop the whole column
Missing Value Imputation
PWD_UPDT_TS
Null percentage = 22.3%
It's a date value, and very hard to fill the null value, therefore, add a column called "ismissing_PWD_UPDT_TS" indicate if PWD_UPDT_TS is missing. If missing, equals 1, else equals 0
CARR_NAME, RGN_NAME, STATE_PRVNC_TXT
Null percentage = 19.35%
It is categorical value and all three missing at the same time, therefore, add the null value to "unknown" category
DVC_TYPE_TXT & AUTHC_SCNDRY_STAT_TXT
DVC_TYPE_TXT Null percent = 12.6%
I chose to use KNN classifier to inpute the null value for DVC_TYPE_TXT and AUTHC_SCBDRY_STAT_TXT because I found there are some association between ALERT_TRGR_CD, DVC_TYPE_TXT, AUTHC_PRIM_TYPE_CD, and AUTHC_SCNDRY_STAT_TXT
Lable encoding
KNN imputation
Replace original null value with predicted value
CUST_STATE
Null percentage = 0.26%
Directly drop the rows with null values
PH_NUM_UPDT_TS
Null percentage = 50.6%
Drop the column
Handling Date&Time
Split the data according to "/" or "-", and add new year, month, day, hour, minute, second related column
Exploring Data Analysis
Numerical Data
Categorical Data
Date&Time data
Data Processing Conclusion
ACTN_CD, ACTN_UNTNL_TXT, TRAN_TYPE_CD only contain single value, will drop all of them
CUST_ZIP and CUST_STATE are perfectly correlated, will drop CUST_ZIP
TRAN_DT contains year, month, day of TRAN_TS, but not hour, minute, second. Will drop TRAN_DT
Drop ACTVY_DT and all related columns
Drop PWD_UPDT_TS and all related columns
Drop CUST_STATS
Drop PH_NUM_UPDT_TS
Data Processing
Handling missing values as in Imputaion step
Split year, month, day, hour,minute, second of each datetime feature into different columns
Drop useless features as suggested in data processing conclusion
Change year, month, day, hour, minute, second features to numerical value
Change target value in training dataset to binary format for modeling
Encoding
Binary Encoding
CARR_NAME and STATE_PRVNC_TXT have too many distinct categories. One-hot-encoding will generate extremely sparse data. Therefore, choose binary encoding
One-hot-encoding
Rest of categorical data will perform one-hot-encoding
Model Training & Selection
Training
Ridge&Lasso Regression
Logistic Regression
Dummy Classifier
KNN
SVM
Decision Tree
Random Forest
Gradient Boosting
AdaBoosting
XGBoosting
Selecting
Select XGBoost since it has best score
Parameter Tunning
learning_rate
max_depth
min_child_weight
subsample
colsample_bytree
n_estimators
Result Prediction